Image and Vision Computing 32 (2014) 348–362

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Ensemble of multiple instance classifiers for image re-ranking ☆

Fadime Sener [a], Nazli Ikizler-Cinbis [b,*]

[a] Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey
[b] Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 14 June 2013
Received in revised form 6 February 2014
Accepted 21 February 2014
Available online 12 March 2014

Keywords:
Image retrieval
Image re-ranking
Multiple Instance Learning

ABSTRACT

Text-based image retrieval may perform poorly due to the irrelevant and/or incomplete text surrounding the images in the web pages. In such situations, visual content of the images can be leveraged to improve the image ranking performance. In this paper, we look into this problem of image re-ranking and propose a system that automatically constructs multiple candidate "multi-instance bags (MI-bags)", which are likely to contain relevant images. These automatically constructed bags are then utilized by ensembles of Multiple Instance Learning (MIL) classifiers and the images are re-ranked according to the final classification responses. Our method is unsupervised in the sense that, the only input to the system is the text query itself, without any user feedback or annotation. The experimental results demonstrate that constructing multiple instance bags based on the retrieval order and utilizing ensembles of MIL classifiers greatly enhance the retrieval performance, achieving on par or better results compared to the state-of-the-art.

1. Introduction

In recent years, there has been an enormous increase in the amount of data stored on the Web, where an important portion of this data is images. Retrieving relevant images according to text-based queries has therefore become an important need. However, text-based image search may perform poorly; the retrieval results are seriously affected by various factors, such as irrelevant or incomplete text surrounding the images, polysemy or synonymy of textual descriptions, and more. Since most of the current search engines (such as Google or Yahoo Image Search) make use of such surrounding textual data, the performance of image retrieval can be relatively lower than expected.

In order to increase the performance of such text-based image retrieval systems, approaches on visual re-ranking have been proposed in recent years. In visual re-ranking, the idea is to explore the initial list of returned images by analyzing their visual content and to propose a new ranking in which more relevant images are ranked higher. Such methods are also referred as relevance-based re-ranking methods [1].

In this paper, we propose such a re-ranking framework that analyses the visual content of the images returned by text-based search engines and improve image retrieval results, by building candidate bags that are utilized by multiple instance classifiers. Our proposed system is unsupervised, in the sense that, it does not need any explicit manual labeling of the images or any user feedback. The only input is a text query, and by evaluating the visual content retrieved by this query, our approach first automatically builds multiple classifiers and then re-ranks the images based on the outputs of these classifiers.

The main idea of the proposed method is to automatically create "*bags*" that will be used with Multiple Instance Learning (MIL). In MIL, the classification is built upon bags as opposed to single instances. In this respect, Multiple Instance Learning is inherently suitable for retrieval problems, since in retrieval, the relevancy of the retrieved images is unknown. We claim that, by using the initial retrieval order of images, we can intelligently build candidate bags that can be used within the MIL framework. The MI-classifiers can then learn the hidden patterns that are common to those images in these candidate bags. Based on the resulting classifiers, the images can be re-ranked so that query-relevant images are ranked higher.

The bag construction step is the key point of the proposed approach. We propose three different ways for building candidate bags, namely dynamic, sliding window and dynamic-sliding approaches. The constructed candidate bags are then used in building multi-instance classifiers. Our algorithm operates on multiple-sized candidate bags, and train classifiers using the visual features extracted from each of the constructed set of bags. An ensemble of MI-classifiers is then formed and the images are re-ranked based on the response of this ensemble. The proposed framework is illustrated in Fig. 1. It is important to note that our aim in this paper is not to introduce a novel ensemble learning method as in [30,31], but to show that with a simple ensemble of MI-classifiers that is only based on visual content of the retrieved images and without using any user feedback, we are able to achieve quite successful re-ranking of the images.

☆ This paper has been recommended for acceptance by Nicu Sebe.

* Corresponding author at: Hacettepe University Department of Computer Engineering, 06800, Ankara, Turkey.
E-mail addresses: fadime.sener@cs.bilkent.edu.tr (F. Sener), nazli@cs.hacettepe.edu.tr (N. Ikizler-Cinbis).

http://dx.doi.org/10.1016/j.imavis.2014.02.014
0262-8856/

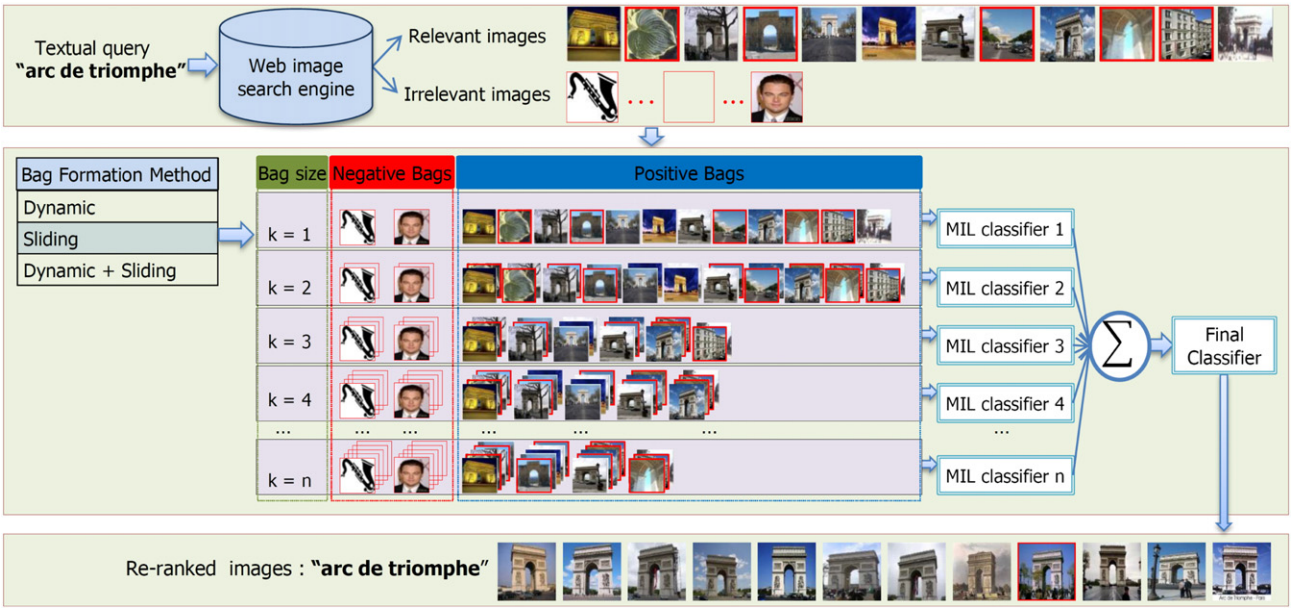

Fig. 1. Our proposed framework for image re-ranking. First, a text query is entered to a web image search engine. Then, multiple size bags are constructed over the initial retrieval order using one of the proposed bag formation methods. Multiple MI-classifiers are learnt using these bags and consequently, the resultant ensemble classifier is utilized for re-ranking the images.

We test our algorithm in Google [2] and Web Queries [3] datasets. The results show that by simply using multiple candidate bags and Multiple Instance Learning in conjunction, our algorithm can perform on par with or better than the state-of-the-art.

The rest of the paper is organized as follows: In Section 2, we review the related literature over the subject. Section 3 introduces the proposed approach of constructing bags for multiple instance classifiers. Experimental evaluation is provided in Sections 4 and 5 we present our conclusions and discussions over the subject with possible future directions.

2. Related work

In this work, we focus on text-based image retrieval and unsupervised re-ranking of images. Our system is based on visual features only; neither additional features, such as textual features, nor auxiliary data, such as user click or feedback data is being used. Our proposed framework relies on multiple bag construction and the use of ensembles of weakly supervised MI-classifiers. Below, we review the related literature, based on the methods and the features in use.

2.1. Image retrieval

In general, image retrieval studies are focused around two main domains; these are content-based image retrieval and text-based image retrieval. Content-based retrieval relies on user-provided query images, where given a query image, visually similar images are searched. An extensive survey on content-based image retrieval can be found in [4]. In text-based image retrieval, on the other hand, the user query is provided in terms of text, as opposed to query images. The aim is to generate a good ranking of the images based on their relevancy to the queried textual term(s).

Initial text-based image retrieval efforts consist of applying text-retrieval techniques to a set of manual annotations that are provided for each image. Providing these manual annotations is a very costly procedure; therefore, image retrieval community inclined towards more automatic approaches, and began to benefit from automatic image annotation and relevance feedback mechanisms. Several automatic image annotation techniques have been proposed (some recent examples include [37,34]), where a model for each semantic concept is learned from a set of captioned images, and consequently the learned model is used to output textual annotations that can be used for retrieval.

Systems that involve user interaction, such as relevance feedback mechanisms have also evolved [36,35]. In relevance feedback systems, the user selects a set of images as relevant or non-relevant, and the system reranks the images based on this human feedback. Since our work does not require any user intervention or labeled data to train upon, we omit an extensive review of such systems and refer the interested readers to recent surveys on these topics [40,41].

2.2. Image re-ranking

Image re-ranking has been a recent topic of interest. Tian and Tao [1] provide a recent and extensive review over the subject. Mainly, the proposed approaches so far differ in the type of features (such as textual, high-level visual and low-level visual features), and the type of learning method (such as clustering, classification, etc.) they utilize.

2.2.1. Clustering-based approaches

Studies [32,39,33] first group visually similar images together and rerank images based on their distances to the discovered cluster centers. Hsu et al. [32] use the information bottleneck principle for discovering the best grouping, whereas typicality-based reranking [33] explores the initial ranking as well as cluster membership to select typical pseudo-positive and pseudo-negative examples. Such approaches may suffer from irrelevant clusters that can be formed from irrelevant images. Moreover, the relevant images may be diversified and may not form dense clusters as required. Similar work by Berg and Forsyth [9] tries to solve such issues by introducing a bit of human intervention, by requiring the user to mark each cluster as relevant or non-relevant.

2.2.2. Topic models

A number of methods use probabilistic topic models [2,5] for image reranking. These studies learn the latent topic amongst the retrieved image list and rerank the images based on the probability that the image belongs to the topic. Fritz and Leibe [5] combine clustering methods with topic models to select the compact subspaces of the latent space and filter out noises. These methods provide promising results,

but may fail when the relevant images are not aggregated in the top retrieval results.

2.2.3. Graph-based models

Graph-based models have also been explored. Hsu et al. [6] propose a random walk based formulation over context graphs for reranking. In their influential work, Ying and Baluja [7] apply the famous PageRank algorithm to the visual content exploration of images.

Several recent studies deal with image re-ranking problem by selecting visually dominant images. Studies [8,28] first remove outliers and search for a confident image set. Morioka and Wang [28] propose to find a confident image set based on sparsity and ranking constraints. These confident images are used as reference points that are further used in a kernel-based re-ranking approach. Similarly, Liu et al. [8] use spectral filter to remove outliers and select a confident set, then apply a graph based re-ranking algorithm.

Although the idea of removing outliers is effective, both of these methods stuck in one dominant group to achieve outlier removal. For a text-based image search query, there may be more than one dominant group and methods based on a single confident set may fail to identify images in different dominant groups. In our work, we do not assume a single visually dominant group; if there is more than one, those are explored in the MI-learning framework.

2.2.4. Additional features

Textual features or other auxiliary data has been explored in quite a number of studies for improving the image re-ranking [9,10]. In [11], Shroff et al. used multimodal features such as text, metadata and visual features together to retrieve and build an automatic re-ranking. Geng et al. [12] propose a content-aware ranking system, in which visual cues are incorporated to the ranking learning process and jointly utilize the textual and visual features. A recent work by Jain and Varma [38] explores user-click data, as well as visual and textual features. Another recent work [29] proposes an algorithm based on deep contexts extracted from textual information surrounding each image. In this work, additional text queries are formed based on the textual context of an image and these queries are used for computing the irrelevancy of an image.

In our approach, we do not make use of any textual cues or any other external source of information such as user-click data, we just use the initial ranking produced by the text query and explore the visual content.

2.3. Multiple Instance Learning (MIL)

MIL methods [13–15] have large applicability in computer vision problems, especially in the cases where manual annotations are expensive or difficult to obtain. This weakly supervised learning paradigm has been used in a wide range of applications, such as object recognition and detection [16,17], tracking [18,19], image classification [20,21], scene classification [15] and more. In this work, we adopt MIL techniques for the problem of image re-ranking.

The work of Li et al. [22,23], which also makes use of MIL for image re-ranking, is the closest to our work, in the sense that they also apply Multiple Instance Learning to image re-ranking. In their framework, they assume that at least a certain portion of a positive bag is of positive instances, and devise a new MIL approach to work over such *constrained bags*. Our proposed framework does not rely on any assumptions about the MI-bags and the positivity of the instances, and does not pose any constraints on the amount of positive instances in a bag. We use the standard definition of MI-learning and therefore, any MI-learning algorithm can be adopted in our system. In the experimental section, we compare our method to Li et al.'s work.

3. Image re-ranking with ensemble of MIL classifiers

We propose a system which automatically learns the queried textual concept by exploring the visual content of the noisy set of retrieved images and produces an improved ranking result. Our formulation is based on multiple instance classifiers, which treat the retrieved images as bags of positive instances. The formation of the "*multi-instance bags* (*MI-bags*)" is the key aspect of our algorithm. During this formation, we do not use any manual labeling of the retrieved images, but only assume that the retrieved set of images include some relevant images.

In this study, we propose a number of methods for constructing candidate bags, so that multiple-instance classifiers learned upon them form discriminative classifiers. These classifiers can then be used for image re-ranking and consequently improve image retrieval performance.

We first review Multiple Instance Learning (MIL) paradigm and discuss why it is suitable for the problem of image re-ranking and categorization. Then, we present our approach on constructing MI-bags for MIL classification.

3.1. Overview of Multiple Instance Learning

In image retrieval, once the text query is input to a text-based image search engine, such as Google or Yahoo Image Search, a set of images is returned. These returned results are not always perfect, and most of the time, irrelevant images occur in higher ranks on the retrieved list. By analyzing the visual content of retrieved images, classifiers for the queried concept can be learned, and using these classifiers the relevant images can be ranked higher in an updated retrieval result.

Working on single image instances and building supervised classifiers using each image would require the availability of user feedback data or large scale annotation effort. When there is no such data available, which is the case with the traditional text-based query system, the text-based retrieval order can provide an initial cue on the relevancy

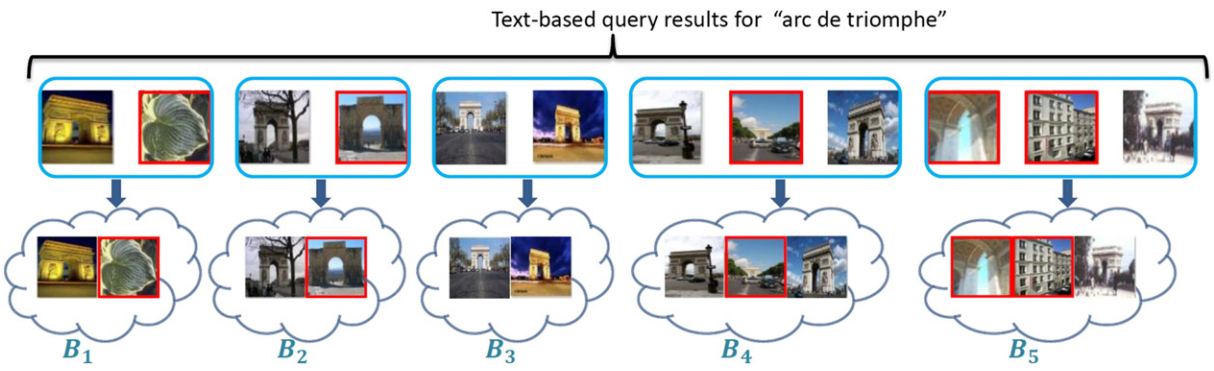

Fig. 2. Formation of dynamic-size bags from the retrieved images. For the images that returned earlier in the list, smaller bags are formed, and for the images that return later in the list, larger bags are formed. In this example, the initial k is 2; for the lower ranks of the text-based retrieval order k value is incremented by 1 and larger bags are formed.

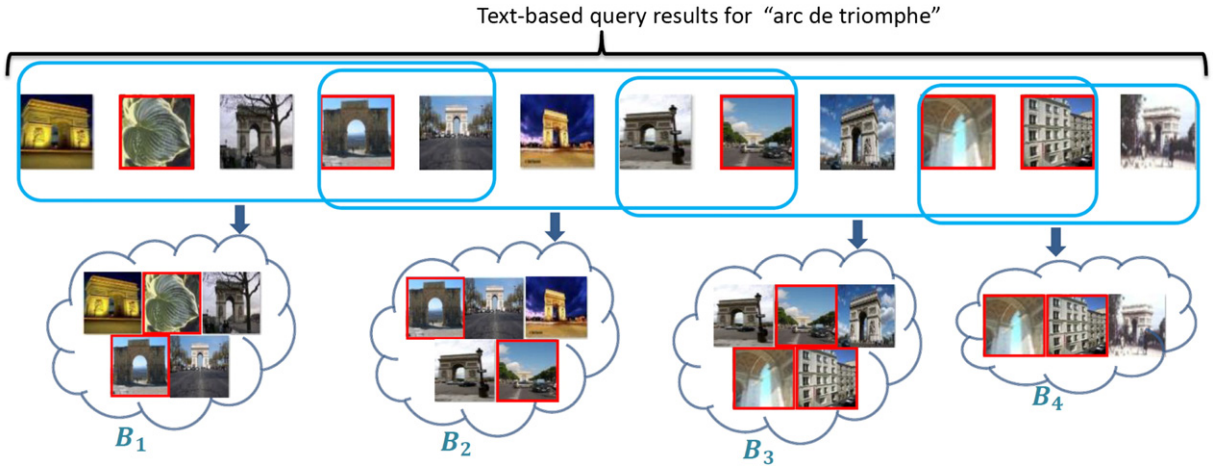

Fig. 3. Sliding window approach for formation of fixed-size bags from the retrieved images. Here k is fixed ($k = 5$) and step size $M = ceil(k/2)$. Sliding window approach generates multiple overlapping bags and provides a dense sampling of the possible bag candidates for MI learning.

of the images to the queried concept. Text-based retrieval order is mostly formed using textual information surrounding the images, user click data, etc., and is likely to contain a certain number of in-class images. Based on this observation, we can assume that in-class images are returned throughout the retrieved list, although these in-class images can be ranked lower in the list or scattered throughout the list.

Since the exact labels for the class of the individual images are unknown, working over single images using supervised classification methods is not possible. However, if we assume that the in-class images are present throughout the list, we can form "*bags*" of the images and assume that each bag contains at least one positive example for the query. By this way, we can utilize Multiple Instance Learning over bags of images.

As opposed to traditional supervised learning, where the learning procedure works over instances x_i and their corresponding labels y_i, Multiple Instance Learning operates over bags of instances, where each bag B_i is composed of multiple instances x_{ij}. This form of learning is referred as "semi-supervised" (or "weakly supervised"), since the labels for the individual instances are not available, and only labels for the bags are given. A bag B_i is labeled as positive, if at least one of the instances x_{ij} within the bag is known to be positive, whereas it is labeled as negative, if all the instances are known to be negative.

As discussed above, Multiple Instance Learning is particularly suitable for our problem. Multiple candidate positive bags can be formed by using the text-based retrieval order of the images and thereon, Multiple Instance Learning classifiers can be used to learn the queried concept.

A problem with the static and non-overlapping construction of the bags (as in [22]) is that the positivity assumption of the bags may not necessarily hold. From the nature of the image retrieval, we can assume that some of the bags contain positive images which are related to the queried concept. However, since we do not use explicit user feedback data, we do not know exactly which bags are indeed positive and which bags are negative in training. In order to deal with this issue, we generate multiple hypotheses for candidate bags from the ordered set of retrieved images and learn multiple MIL classifiers over each hypothesis. Our approach then combines multiple classifiers and re-ranks the images based on their classification scores.

3.2. Constructing candidate bags

Candidate bag generation is the key aspect of our approach. We evaluate different ways for constructing candidate multiple instance bags (MI-bags) which will be used in learning multiple instance classifiers. These different schemes are namely fixed-size bags, dynamic-size bags, sliding window and dynamic-sliding approaches. We now describe each of these approaches in detail.

3.2.1. Fixed-size bag construction

The simplest way to build candidate bags for employing Multiple Instance Learning is to use fixed-size bags. In this approach, the initial list of images is divided into small subsets, i.e. bags, in which each bag contains k images. Then, these bags are utilized in MIL setting as positive instance bags. This approach is similar to the initial bag formation of [22], with the exception that there is no random subset selection from the initial retrieval order.

More formally, given ranking R, the set of retrieved images is divided into equal k-sized bags, so that each bag contains k images based on R. In

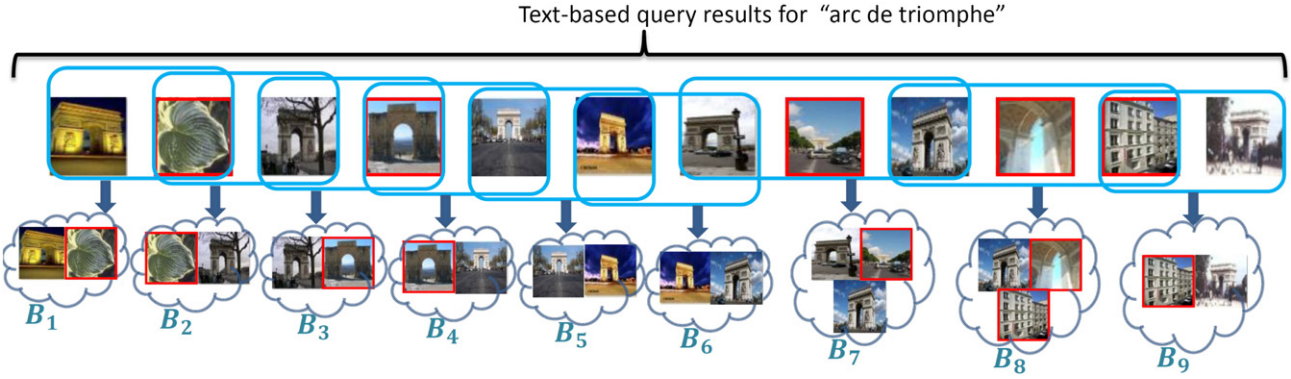

Fig. 4. Dynamic-sliding window approach for formation of candidate bags from the retrieved images. Here k is dynamic (initial $k = 2$ and increase rate $\sigma = 1$) and step size is $M = ceil(k/2)$.

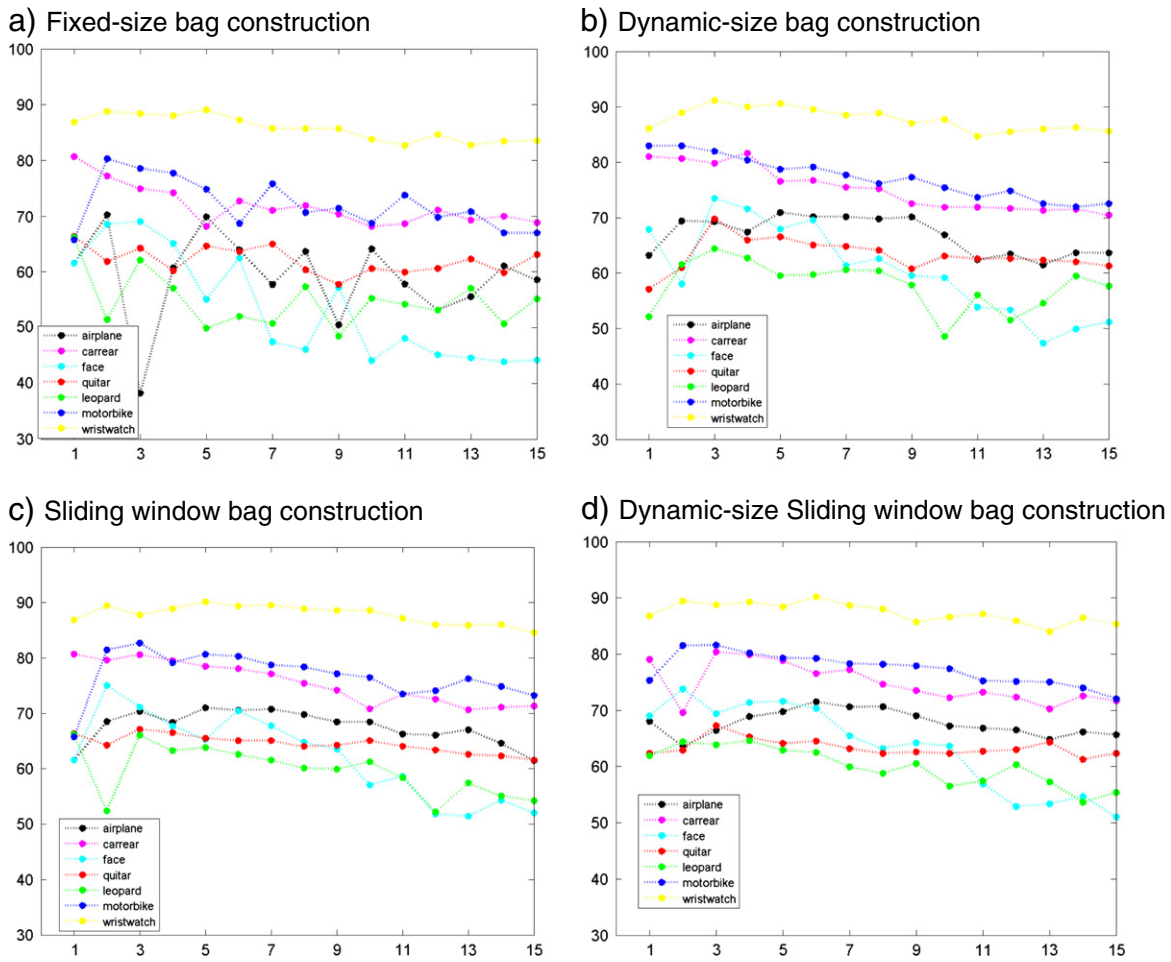

Fig. 5. Effect of choosing different bag sizes k using the four proposed MI-bag construction methods. The results presented here are the average precision (AP) values achieved on the Fergus dataset [2]. We observe that the fixed-size bags are affected very much from the choice of k and produces rather unstable results, whereas the sliding window (SW) and dynamic-size sliding window (DSW) approaches are less affected from the change in k. From this figure, we also observe that there is no global optimal choice of k that produces the best results for all the queries.

this construction phase, first k images that have ranks r_1 to r_k are assigned to bag B_1, images from r_{k+1} to r_{2k} are assigned to bag B_2 and so on.

In the Experiments section, we present results with different k values, and see how the choice of k affects image retrieval performance. Since we do not have an explicit information on the positivity of the retrieved images, the best choice for k can be determined empirically. However this would require the availability of manually labeled set of images. In order to overcome this issue, we generate multiple candidate bags with varying k, and train classifiers using each of the constructed set of bags. Using the ensemble of these classifiers, we utilize the outputs of multiple candidate bags of varying sizes, thus bypass the selection of the optimal k value. This approach is further discussed in Section 3.3.2.

3.2.2. Dynamic-size bag construction

As discussed in the introduction, text-based search engines use surrounding text information accompanying images to retrieve relevant image data. While this text information is mostly noisy and incomplete, it can be seen as an initial point of reference for evaluating the images. In this context, we observe that, while the image search engine performance is far from perfect, the images returned earlier in search ranking, tend to be more relevant to the queried concept. Based on this observation, in order to increase the likelihood of each bag to contain an in-class image, we can form relatively smaller bags for the top ranks of the retrieved list and relatively larger bags from the lower ranks of the list. We call this procedure "*dynamic-size bags*".

Assuming that the relevancy of the images decreases as the rank of the image increases, we can increase the bag size gradually at each γ interval of received images. More formally, given ranking $R = r_1 \ldots r_N$, where N is the size of the image set, the set of retrieved images that have ranks r_1 to r_γ are divided into k-sized bags, images with ranks $r_{\gamma+1}$ to $r_{2\gamma}$ are divided into $(k+\sigma)$-sized bags, where k is the initial bag size, and σ is the amount of size increment. This procedure is illustrated in Fig. 2.

By this way, since the images returned later in text-based search ranking tend to be less relevant than the images returned earlier in the search, by increasing the bag size, the probability for each positive bag to include a positive instance is likely to be increased. In the Experiments section, we evaluate how varying k, γ and σ affect the retrieval performance.

3.2.3. Sliding window bag construction

Since the retrieved images do not have explicit labels, we cannot make sure that the candidate positive bags indeed include a positive instance for the MIL training. In order to deal with this issue, we can generate multiple overlapping bags. By following a sliding window approach, we can generate multiple bags, where at least a portion of these bags are assured to include positive instances. By dense sampling of bags in this way, we make sure that a large portion of the possible bag combinations are evaluated.

Fig. 6. Example retrieval list and MI-bags constructed via fixed-size bag construction method using $k = 3$. Each $k = 3$ images form a bag, and the mark on upper corner shows whether the bag is an actual positive. With fixed size bag construction, the optimal k for each query can be varying.

The sliding window procedure for building bags is shown in Fig. 3. This approach is analogous to the sliding window approach for object detection, where a window is slid over an image to search for particular occurrences of an object. In our context, by sliding a window over the sets of image instances, we consider each set of instances that falls within the same window as a candidate bag that will be used in MIL procedure.

More formally, given a ranking R of image set $I = \{i_1, \ldots, i_N\}$, starting from image ranked in R_1, we create a k-size bag where images from $R_1 \ldots R_k$ are assigned to B_1. At each sampling step, we increase the index by step size $M = ceil(k/2)$ and create a new bag, so that each new bag is composed of the images within retrieval rank $\{R_{(i-1+M)} \ldots R_{(i-1+M+k)}\}$.

3.2.4. Dynamic-sliding bag construction

This bag construction procedure is the combination of sliding window and dynamic-size bag construction approaches. In this approach, a window is slid over the initial retrieval list of the image instances, and each set of instances that falls in the same window is taken as a candidate bag. As opposed to using a fixed-size window, the size of the sliding window is gradually increased as the window is moved down the retrieval list.

More formally, given a ranking R of image set $I = \{i_1, \ldots, i_N\}$, starting from image ranked in R_1, we create a k-size bag where images from $R_1 \ldots R_k$ are assigned to B_1. At each sampling step, we increase the index by step size $M = ceil(k/2)$ and create a new bag, so that each new bag is composed of the images within retrieval rank $\{R_{(i-1+M)} \ldots R_{(i-1+M+k)}\}$. In dynamic-sliding procedure, the bag size k is increased gradually with a rate of σ at each γ interval of retrieved images. This process is depicted in Fig. 4.

We evaluate all of these aforementioned bag construction procedures in detail in the Experiments section.

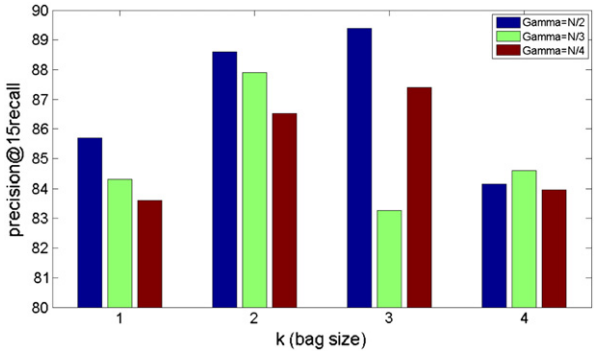

Fig. 7. Mean average precision at 15% recall for the dynamic-size bag construction where $\sigma = 1$ and k changing in Fergus [2] dataset.

3.2.5. Constructing negative bags

In order to use negative bag constraints of Multiple Instance Learning, it must be made sure that the constructed negative bags do not contain any positive instances. For this reason, while constructing negative bags, we use the images returned for queries other than the search query. We apply a similar scheme that sequentially forms the MI-bags based on the order of the images. However, it is possible that for non-relevant queries, some negative image pattern may emerge amongst the retrieved set for negative queries. In order to refrain from such a pattern, we first cluster the images returned for non-relevant queries by using k-means. Then, the cluster center order is randomized and the images are re-ordered based on the distances to these cluster centers. Then, this new order is used as the negative image set order. By this way, it is made sure that the order of images is randomized and the similar images are not scattered through the list of negative images, to avoid misleading patterns. Once the randomized list of negative images are established, we form fixed-sized bags over this negative image set.

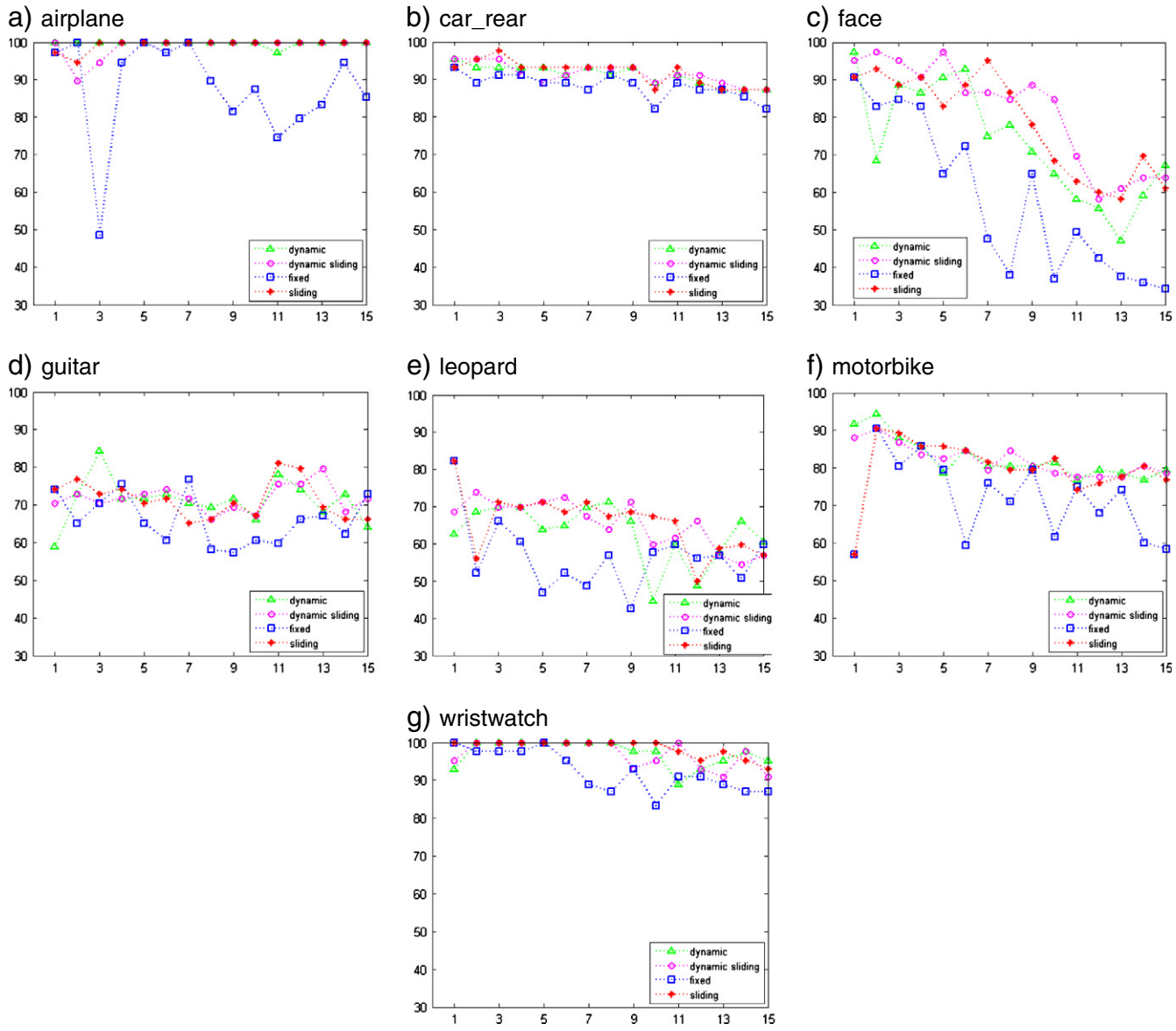

Fig. 8. The effect of choosing different bag sizes with different bag construction approaches and varying initial bag sizes k on the Fergus Google dataset [2]. Here, the precisions at 15% recall level are shown.

3.3. Classification

Once the positive and negative bags are formed via one of the proposed schemes, Multiple Instance Learning algorithms can be applied using the constructed MI-bags. We now present the details of this classification stage.

Table 1
Precision at 15% recall for the dynamic-size bag construction where $\gamma = N/2$ and $k = 2$ for the first interval. The highest precision is shown in bold.

σ	Airplane	Car_rear	Face	Guitar	Leopard	Motorbike	Wrist watch	Mean
1	100	93.18	88.64	72.88	71.14	94.29	100	**88.59**
2	100	93.18	65.00	76.79	69.81	91.67	100	85.21
4	89.64	95.35	97.50	69.35	68.52	91.67	100	87.43
6	100	93.18	95.12	60.56	71.15	85.71	100	86.53
8	100	93.18	88.64	72.88	68.52	74.16	97.56	85.00
10	100	97.62	84.78	66.15	71.15	90.41	97.56	86.81

3.3.1. MIL classification

Our MI-bag formation procedure is independent of the choice of the multiple instance classifier, therefore any multiple instance classifier can be used with our framework. In this study, we utilized Multiple Instance Learning with Instance Selection [13] (MILES) algorithm as the MI-classifier. MILES [13] algorithm works by embedding the original feature space x, to the instance domain $\mathbf{m}(B)$. Each bag is represented by its similarity to each of the instances in the dataset. The similarity between bag $\mathbf{B}_i$ and concept c_l is defined as

$$s(c_l, \mathbf{B}_i) = \max_j \exp\left(-\frac{D\left(x_{ij}, c_l\right)}{\sigma}\right), \quad (1)$$

where $D(x_{ij}, c_l)$ measures the distance between a concept instance c_l and a bag instance x_{ij} and σ is the bandwidth parameter. For $D(.)$, any standard distance measure that is suitable for the feature space can be used. In our case, since all the features are histogram-based, we can use the χ^2

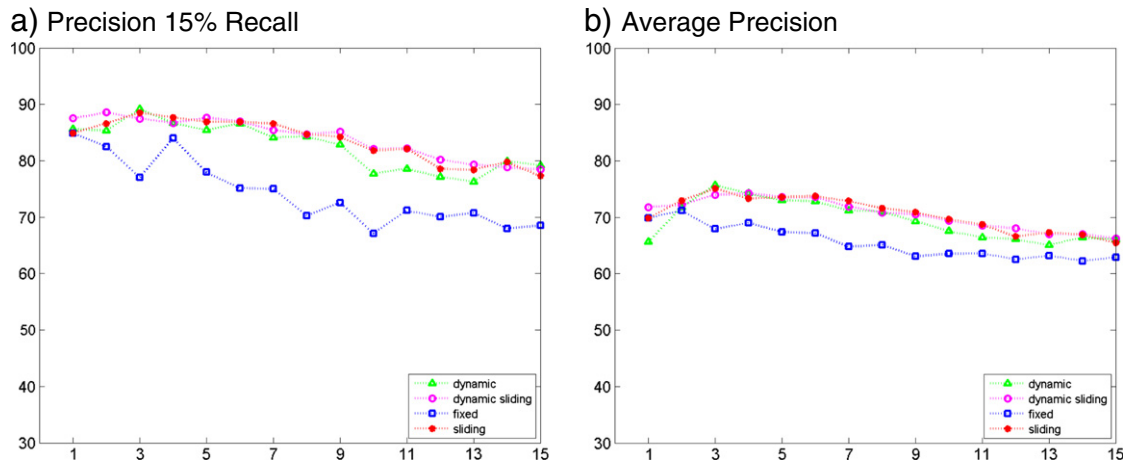

Fig. 9. Mean performance of the four different MI-bag construction methods on the Fergus Google dataset [2] with respect to changing bag size k. To the left, the precisions at recall 15% are shown, and to the right, the average precision values are given. Sliding window (SW) based MI-bag construction methods are more likely to produce better results.

distance $D\left(x_{ij}, c_l\right) = \chi^2\left(x_{ij}, c_l\right) = \frac{1}{2}\sum_d \frac{\left(x_{ij}(d) - c_l(d)\right)^2}{x_{ij}(d) + c_l(d)}$, where d is a feature dimension of the instance feature vector. We evaluate the effect of choosing different distance functions in the experimental evaluation.

Each bag can then be represented in terms of its similarities to each of these target concepts and this mapped representation $\mathbf{m}(B_i)$ can be written as

$$\mathbf{m}(B_i) = [s(c_1, B_i), s(c_2, B_i), \ldots, s(c_N, B_i)]^T. \quad (2)$$

We then use an SVM classifier over this embedded representation. The original MILES formulation incorporates an L1-regularized linear SVM, which enforces some sparsity on the data. In our case, since the retrieval data can have multiple modes, we experience that using L2-regularized SVM is better suited for this purpose.

3.3.2. Ensemble of MIL classifiers

While forming the positive bags for the MIL framework, the most crucial parameter is the bag size k. The optimal k depends mostly on the order of initial retrieval. Since our algorithm does not make use of any explicit user feedback or labeled data, determining the optimal k value that is generic and optimal for each query is not possible.

We have empirically observed that the performance is largely dependent on the selection of k parameter and the optimal choice of k is largely query dependent. Since there is no strongly supervised training set by definition of the reranking problem, it is difficult to reliably optimize k in a query-specific manner. To overcome this issue, we learn an ensemble of MI classifiers, each of which works on multiple bags formed using different k values. The final ranking is obtained by model averaging, where the score of a retrieved image is the average of all MI classifier responses.

Our simple yet effective ensemble scheme not only bypasses the problem of choosing k, but also results in stronger classifiers. We have observed that ensemble based reranking typically outperforms any particular choice for the k parameter. In Section 4, we experimentally evaluate ranking performance based on particular k values and ensemble of MIL classifiers.

4. Experiments

In this section, we evaluate the proposed MI bag construction approach and ensemble classification.

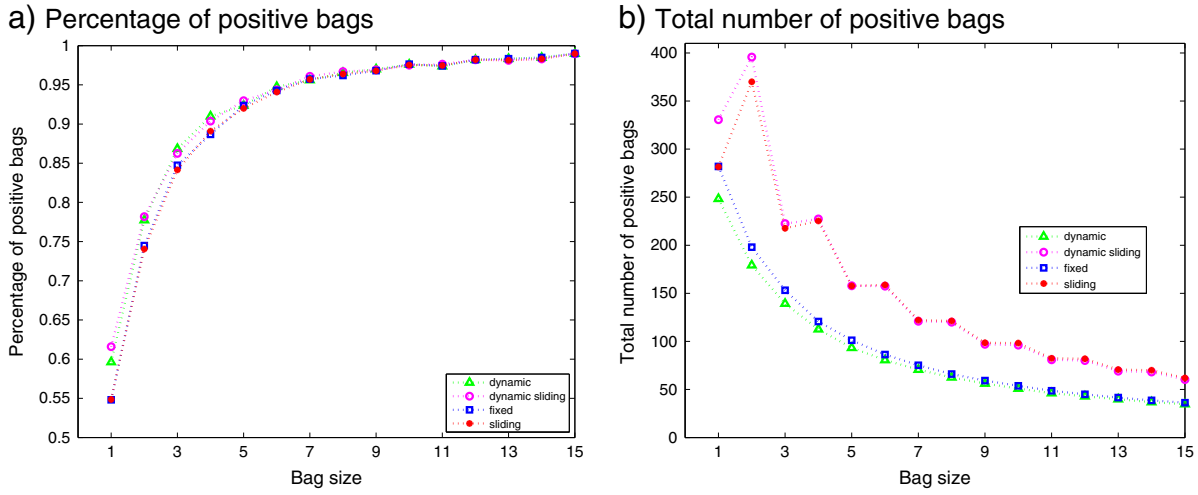

Fig. 10. How many of the constructed bags are indeed positive? Number and percentage of positive bags constructed as the result of the four different MI-bag construction methods on the Fergus Google dataset [2]. While for larger k, percentage of positive bags increases, the number of positive bags gets considerably smaller, which reduces the effectiveness of the MI-learning phase.

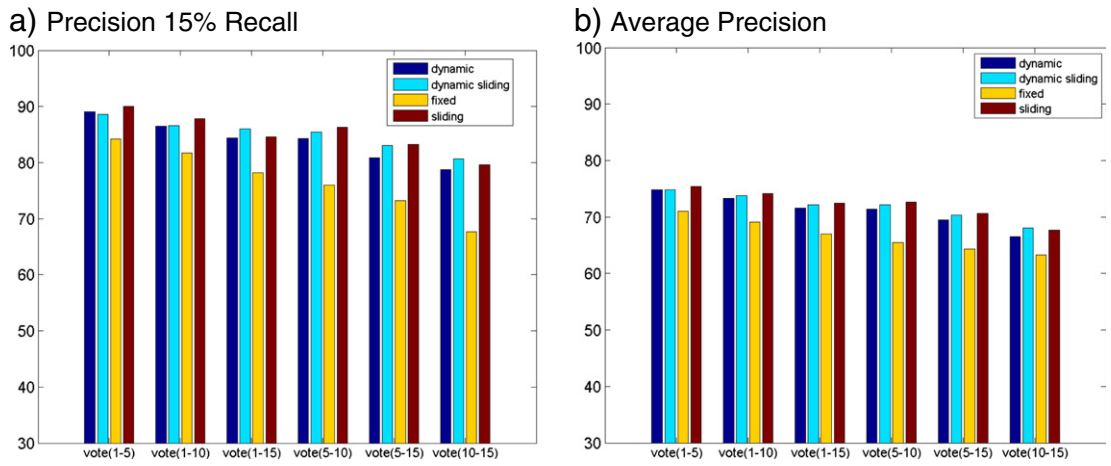

Fig. 11. Using ensemble of MI classifiers with different bag sizes k and different bag construction schemes over Google dataset. Vote (k_1, k_2) shows that $k \in k_1 \ldots k_2$. In this dataset, using sliding window (SW) with fixed size bags produces the best result, whereas using SW with dynamic size windows is the second best. According to these results, using classifiers with bags built with $k \in 1 \ldots 5$ gives the highest precision.

4.1. Datasets

In order to evaluate the performance of our method, we use two benchmark datasets. The first is the Fergus dataset [2] and the second is the Web Queries [3] dataset.

Fergus Google dataset [2] has been collected via text queries from the Google Image Search. This dataset consists of 7 categories (airplane, cars rear, face, guitar, leopard, motorbike, and wrist watch) and each of these categories includes about 600 images on average. For each category, labeling is done with 0 = "Junk", 1 = "Intermediate" and 2 = "Good" for each image. On average there are 30% "Good" images without major occlusion, but no constraints on viewpoints, scaling and orientations, 20% "Intermediate" images have lower quality when compared to "Good" images, have extensive occlusion and image noise, and 50% "Junk" images that are irrelevant to the category.

Web Queries [3] is a recently compiled dataset, which includes 353 web image search queries. These queries are selected among the frequent terms submitted to image search engines. There are more than 200 images for 80% of the images, and the dataset has 71,478 images in total. The images have been scaled to fit within a 150×150 square, keeping the original aspect ratio. Some example topics in this dataset are maps, animals, celebrities from TV, flags, logos, buildings, and so on.

4.2. Feature extraction

To capture the visual content, each image is represented via its bag-of-words (BoW) histograms. First, dense SIFT descriptors [24] are extracted from each image using VLFeat library [25]. We then cluster these descriptors using k-means (where we set $k = 1000$ in our experiments) and form the visual codebook. Then, each image is represented with its histogram of codewords. While forming the image representation, 2×2 spatial tiling is applied to account for coarse spatial information. Each of the local spatial histogram is concatenated with the global BoW histogram of the whole image. The resulting feature vector size is therefore 5000 (1000 for the overall image histogram, 1000 for each spatial quadrant).

4.3. Evaluation of the bag-size and bag construction approaches

We first investigate whether there is a fixed bag size k that produces effective results for each dataset. Extensive evaluation of choosing the bag-size k and different MI-bag construction approaches over the Google dataset [2] are given in Figs. 5 and 8. Below, we describe each of the experiments in greater detail.

4.3.1. Fixed-size bag construction

We first evaluate the simplest bag construction method, i.e. using fixed-size bags. For each category, we show the effect of using various bag sizes $k = 1, 2, \ldots, 15$ in terms of average precision (AP) in Fig. 5(a). The results show that fixed-size bag construction is quite dependent on the choice of k. We observe that the average precision is mostly higher for the lower values of k (such as $k = 1, \ldots, 3$), however, there is no optimal value which performs best for each of the categories. Moreover, the performance fluctuates quite rapidly based on the choice of k. This is not surprising, since for each image query, the relevancy of the initial retrieved ranking list is quite versatile and dependent on many factors of used text-based retrieval scheme. Example initial ranking lists and fixed-size bags formed with $k = 3$ can be seen in Fig. 6. We see that for some choice of k, the re-ranking performance increases, this is due to the generation of more suited MI-bags. On the contrary, for some choice of k, the performance decreases and this is due to the increased noise content in the MI-bags or decreased number of bags that is used as an input to the MIL algorithm. Since there are no explicit

Table 2
Precision at 15% recall level is shown. D corresponds to the distance function used in Eq. (1) for MIL instance embedding step, and BoW representations are used either in standard or in Hellinger-kernelized form.

D	BoW	Airplane	Car_rear	Face	Guitar	Leopard	Motorbike	Wrist watch	Mean
euc	normal	100	95.35	95.12	78.18	71.15	90.41	100	90.0
euc	Hellinger	100	97.62	100	89.58	62.71	95.65	100	92.2
chi	Normal	100	100	97.5	82.69	75.51	97.06	100	**93.3**
chi	Hellinger	100	100	92.86	84.31	67.27	95.65	100	91.4

Table 3
Average Precision: Parameter optimization and best method. ed: Euclidean distance for MILES, ed-sqrt: Euclidean distance for MILES with sqrt of BoW histograms, chi: chi-square distance for MILES, and chi-sqrt: chi-square distance for MILES with sqrt of BoW histograms.

D	BoW	Airplane	Car_rear	Face	Guitar	Leopard	Motorbike	Wrist watch	Mean
euc	normal	71.56	80.78	70.31	66.56	64.58	83.05	90.75	75.38
euc	Hellinger	68.19	82.22	74.39	71.81	60.62	86.56	92.16	76.56
chi	normal	68.40	83.03	73.49	72.02	64.10	87.00	92.72	77.25
chi	Hellinger	72.11	83.05	73.04	73.04	61.81	85.46	92.95	**77.35**

labels or user feedback, it is not possible to select the optimal k for each query.

4.3.2. Dynamic-size bag construction

In dynamic bag construction, we divide the retrieved list of N images to subsets of size $N/2$ and for each subset, the size of the MI-bag is incremented by 1 (i.e. $\gamma = N/2$ and $\sigma = 1$). Fig. 5(b) shows the performance of this method using varying k. In this figure, as with the case of fixed-size bags, the performance is highly sensitive to the choice of k. However, especially for some values of k, the results are better than using fixed-size bags. This result is in accordance with our initial observation that the retrieved list of images tends to contain relevant images ranked higher in the list, whereas the lower portions of the retrieval list contain images that are less relevant. Since the frequency of seeing relevant images decreases as we move down the list, increasing the MI-bag size affects the performance positively.

For dynamic-size bag construction, we evaluate the choice of σ (amount of increase in each subinterval) and γ (the interval size). The results are given in Table 1 and in Fig. 7, respectively. In Table 1, we look into the effect of increasing the bag sizes as we move further down the initial retrieval list. As these results show, in our experiments, we observe no significant trend related to the choice of σ. Overall, increasing the bag size is more effective compared to using fixed-size bags, whereas using gradual increments is likely to be more promising. Based on this observation, we set $\sigma = 1$ for the rest of the experiments.

In Fig. 7, we show the effect of varying γ intervals, where the retrieval list is divided into $N/2$, $N/3$ and $N/4$ intervals and in each interval the bag size is incremented by 1. We observe that, $\gamma = N/2$ produces slightly better results, thus set $\gamma = N/2$ for the rest of the experiments.

4.3.3. Sliding window bag construction

Sliding window (SW) approach for constructing MI-bags can be used with both fixed-size bags and dynamic-size bags. For the case with the fixed-size bags, the results are given in Fig. 5(c). From this figure, we observe that SW approach is less affected from the choice of k compared to fixed-size or dynamic-size bag construction methods. On the other hand, still, there is no global k that is optimal for every query. In Fig. 5(d), the results when sliding window approach is used with dynamic-size (dynamic-SW) bags are presented. We observe a similar trend in these results.

Fig. 8 compares the performance of all the four bag construction methods on different queries in Google [2] dataset. As it can be seen, amongst all four bag construction approaches, the fixed-size bag construction performs the worst. The best performance is achieved by SW approach either with fixed or dynamic-size bags. Fig. 9 shows the mean performance of those methods with respect to varying k. Again, for different choices of k, either SW or dynamic-SW approach performs the best. We also observe that the performance is relatively higher for lower k values. This implies that, as the bag size increases, the amount of noise present in each bag becomes more dominant and this situation affects classification performance in a negative way.

In order to investigate the bag construction process in more detail, we also present the percentage (Fig. 9) and the total number of bags (Fig. 9) that are indeed positive. For small k, the ratio of actual positive bags to the total number of constructed positive bags is also small. As the bag size increases, more of the constructed bags become positive since it is more likely for a large bag to be positive — e.g., if there is only one bag that includes all the retrieved images, the bag will be positive even if the returned images contains only one relevant image. However, in our experiments we observe that our algorithm is much successful using smaller k where $k = 1 \ldots 5$. For small k, the total number of bags, as shown in Fig. 9 is higher, and MILES algorithm benefits from using a large number of positive bags in training.

4.3.4. Using ensembles of MIL classifiers

The results show that the re-ranking performance is quite affected by the choice of k parameter. Choosing the optimal k parameter is not feasible, since our method does not use any supervision or user feedback. In order to deal with this issue, we propose to train multiple MI classifiers that work on bags of varying sizes. Ultimately, the responses of these classifiers are combined for final decision. In this way, we bypass the need of choosing the bag size and reduce the number of parameters that needs to be tuned.

The results of using such ensemble classifiers are shown in Fig. 11. From these results, we observe that combining multiple classifiers produces more effective re-ranking results, and on average, 1% to 5% point precision gain is achieved as opposed to using single MI-classifiers with a particular choice of bag size. The best performing method in Google dataset is using sliding window with fixed-size bags, where the bag size is $k \in 1 \ldots 5$. Using this range seems to perform the best for all methods in our experiments, therefore, we construct multiple bags of size 1 to 5 in the rest of the experiments.

4.3.5. Evaluation of distance function and BoW representation

We further evaluate the effect of the distance function used in instance embedding step of the MILES classifier, i.e. D function in

Table 4
Clustering-based MI-bag construction performance on Fergus [2] is given. Here, precisions (%) at 15% recall rate are reported. The best performance of single sized MI-bags are achieved with $k = 15$. Overall, using ensemble of MI-classifiers that are learned upon bags of different sizes gives better performance.

	Airplane	Car_rear	Face	Guitar	Leopard	Motorbike	Wrist watch	Mean
k = 5	64.8	91.1	86.7	45.8	66.1	94.3	100	78.4
k = 10	77.8	95.4	90.7	47.8	61.7	90.4	95.2	79.8
k = 15	83.3	95.4	88.6	71.7	52.1	85.7	93.1	81.4
k = 20	81.4	93.2	95.1	37.7	60.7	85.7	83.3	76.7
Ensemble	94.6	95.4	100	45.3	60.7	94.3	97.6	83.9

Table 5
Comparison to state-of-the art on Google dataset [2]. In this table, precisions (%) at 15% recall are reported. The "good" and "intermediate" images are treated as positive, whereas the "junk" images are considered as negative. The highest performance is shown in bold.

	Airplane	Car_rear	Face	Guitar	Leopard	Motorbike	Wrist watch	Mean
Google	70.0	69.5	43.8	56.6	66.1	72.5	88.9	66.8
SVM [11]	58.5	N/A	N/A	70.0	49.6	74.8	98.1	70.2
WsMIL [27]	100	81	57	52	66	79	95	75.7
MIL-CPB [23,22]	–	–	–	–	–	–	–	85.6
PMIL [22]	100	75.3	89.9	82.7	86.2	76.6	95.7	86.6
Ours	100	100	97.5	82.7	75.5	97.1	100	**93.3**

Table 6
Comparison to state-of-the art on Google dataset [2]. In this table, precisions (%) at 15% recall are reported. Here, only images with "Good" label are treated as positive, whereas the "intermediate" and "junk" labeled images are considered as negative. The best performance is shown in bold.

	Airplane	Car_rear	Face	Guitar	Leopard	Motor bike	Wrist watch	Mean
Google	50	41	19	31	41	46	70	43
SVM [11]	35	–	–	29	50	63	93	54
LogReg [3]	65	55	72	28	44	49	79	56
WsMIL [27]	*	*	*	*	*	*	*	58.9
[2]	57	77	82	50	59	72	88	69
[8]	86	100	75	58	63	79	100	80
LDA [5]	100	83	100	91	65	97	100	**91**
Ours	100	100	97.5	82.7	47.1	89.8	100	88.2

Eq. (1). The precisions at recall 15% and average precisions are presented in Tables 2 and 3, respectively. The experiments show that when Euclidean distance is used, using the square rooted BoW feature vector, which is equivalent to Hellinger kernel over BoW vectors [26], produces better results. Using chi-square distance with standard BoW representation yields the highest precision value at recall 15%. Note that using chi-square distance with square rooted BoW features yields slightly higher average precision.

4.4. Comparison to state-of-the-art

In order to evaluate our method's performance with respect to the existing approaches in the literature, we first compare our algorithm to the clustering-based bag formation method, since a number of studies [23,5] have benefitted from clustering to find the most dominant pattern amongst the retrieval list. For this evaluation, in a similar setting to [23], we set number of clusters to $m = \lceil (T/k) \rceil$ where $k = 5, 10, 15, 20$ represents the bag size and T is the total number of images for a text query. Using this setting, we applied k-means clustering over the initial retrieval list and use the clusters that includes $\geq k$ images as the MI-bags. Over these bags, we learned MILES classifiers. Finally, we employ ensemble learning using three classifiers obtained for different bag sizes. In our evaluation, we have used the same negative set and best settings that are used in our best method.

The clustering-based results are given in Table 4. The best results are achieved when the $k = 15$, i.e., when the clusters that have 15 or more elements are used as MI-bags. As it can be seen, using ensembles that are formed with different bag sizes outperform using single MI-classifiers, achieving a precision of 83.9 as opposed to 81.3. Our method, on the other hand, achieves a precision of 93.3 on this dataset, significantly outperforming the clustering-based bag formation. This may be due to the small number of bags presented to the MI-learning as the result of clustering. From MIL perspective, clustering many good images together may decrease the applicability of a MI-learner. If all positive instances of a query is clustered together into a single good cluster, then there would be a single bag to train upon. This may reduce the effect of the MIL classifier, as it would look for a consistent pattern between bags. We also observe this case in Fig. 10, when the bag size k is small, i.e. in the presence of more training bags, the performance of our learning framework is higher.

Next, we compare our approach to state-of-the-art approaches both on Fergus and on Web Queries datasets. In Tables 5 and 6, the comparisons for the Fergus dataset are given. In this table, Ours indicates the results ensembles of MI classifiers with $k = 1 \ldots 5$ where the MI-bags are constructed via sliding window (SW) with fixed size bags, since this method performs the best amongst the four alternatives. Chi-square distance is used for MIL instance embedding stage and L2-regularized linear SVMs are used over the embedding space.

In the literature, there are two different evaluation setups for this dataset. In the first setting (results in Table 5), both the "Good" and "Intermediate" images are taken to be positive, and "junk" images are considered to be negative, whereas in the second setting (results in Table 6), only "Good" images are considered to be positive. We believe that, both "Good" and "Intermediate" images should be considered as positive, since "Intermediate" images are also related to the keyword category as described in [2]; they just contain lower quality images with possible occlusions and substantial image noise.

As the results indicate, our method achieves superior performance and is able to identify images of the queried concept in "Intermediate" quality images as well as in "Good" images. This demonstrates that, our method is able to identify queried concept in spite of the noise, low quality or occlusions. In Table 6, when only "Good" images are considered to be positive, the performance is slightly lower; this is probably due to the related patterns discovered in some "Intermediate"-labeled images being ranked higher.

In Web Queries dataset, we also employ ensembles of MIL classifiers learned over multiple bags, constructed by sliding window approach, where $k = 1 \ldots 5$. Euclidean distance is used for MIL embedding stage. In this dataset, since the modalities within the queries are higher,

Table 7
Comparisons to state-of-the art on Web Queries dataset [3] with respect to the mean average precisions (MAPs).

Method	MAP
Search engine	56.99
[3](Visual only)	64.9
BLVS [28]	67.0
[3](Visual + textual)	67.3
Deep contexts [29]	70.5
SpecFilter + MRank [8]	**73.76**
Ours	71.08

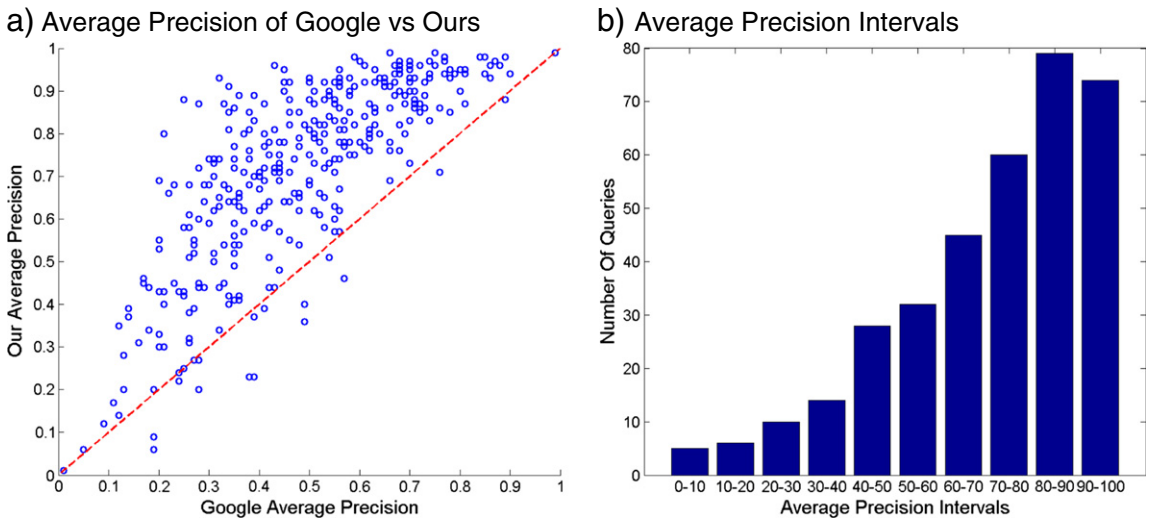

Fig. 12. a) Comparison of our method wrt search engine in terms of individual query APs in Web Queries dataset. We observe that for most of the queries, our method provides higher APs. b) In the result of our method, the distribution of the query APs are shown. Approximately half of the queries have APs ≥ 80.

SVMs with RBF kernel tend to be more effective. Table 7 shows the overall results. Our method achieves a MAP of 71.08% on this dataset, which is comparable to state-of-the-art.

We further evaluate our method's performance with respect to the initial search engine ranking in Fig. 12. Fig. 7 shows the average precisions (AP) of our re-ranking method as opposed to their counterpart search engine ranking APs. Out of 353 queries of Web Queries dataset, the AP has degraded in only 14 queries when using our re-ranking method, and most of the time, our method provides superior ranking compared to the search engine. For some queries that have APs as low as 0.2 or 0.3 in the initial search engine ranking, our method is able to improve the AP to 0.80 and 0.90. Note that, our method does not make use of any auxiliary data, textual data or explicit detector/classifier; it relies solely on the visual content and the initial ranking of the images. From Fig. 7, we also observe that most of the queries fall into the high precision range, approximately half of the queries have APs greater than 0.8. In Fig. 14, some qualitative examples for the re-ranked retrieval lists are given for the Web Queries dataset. Note that our method is able to successfully re-rank various images of queried concept.

4.4.1. Cases of failure

In order to gain further insight about our method's performance, we look at the individual query performance with respect to the positive instance percentage for the queries. Fig. 13 depicts this evaluation. The linear correlation between the two axes in this graph is rather expected for all methods, since as the percentage of positives increases in the set, the average precision also increases. We observe that our method

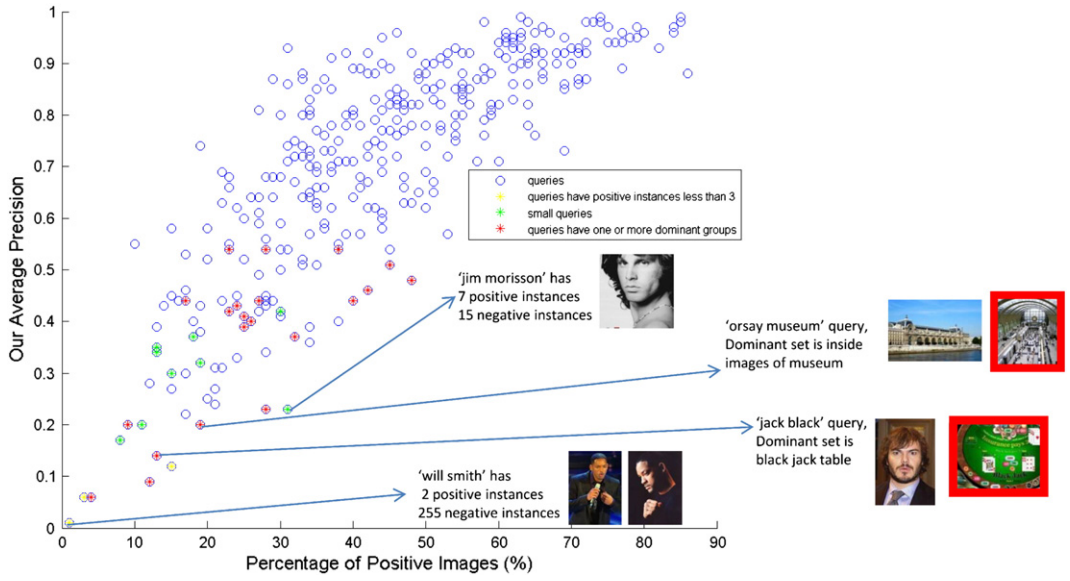

Fig. 13. Our method's average precision vs. the percentage of positive images returned by the search engine. When the number of actual positive instances returned by the initial retrieval is very low for some query (shown in yellow), the classifiers are not able to form reliable models for the queried concepts. Similarly, if the returned image list is relatively sparse, i.e. if it does not include many examples, the AP can also be low (queries shown in green). Another interesting observation is that, when the queries include more than one dominant set (shown in red), the multiple instance learners can focus on the unintended dominant set, and as a result, the re-ranked list can have a lower AP.

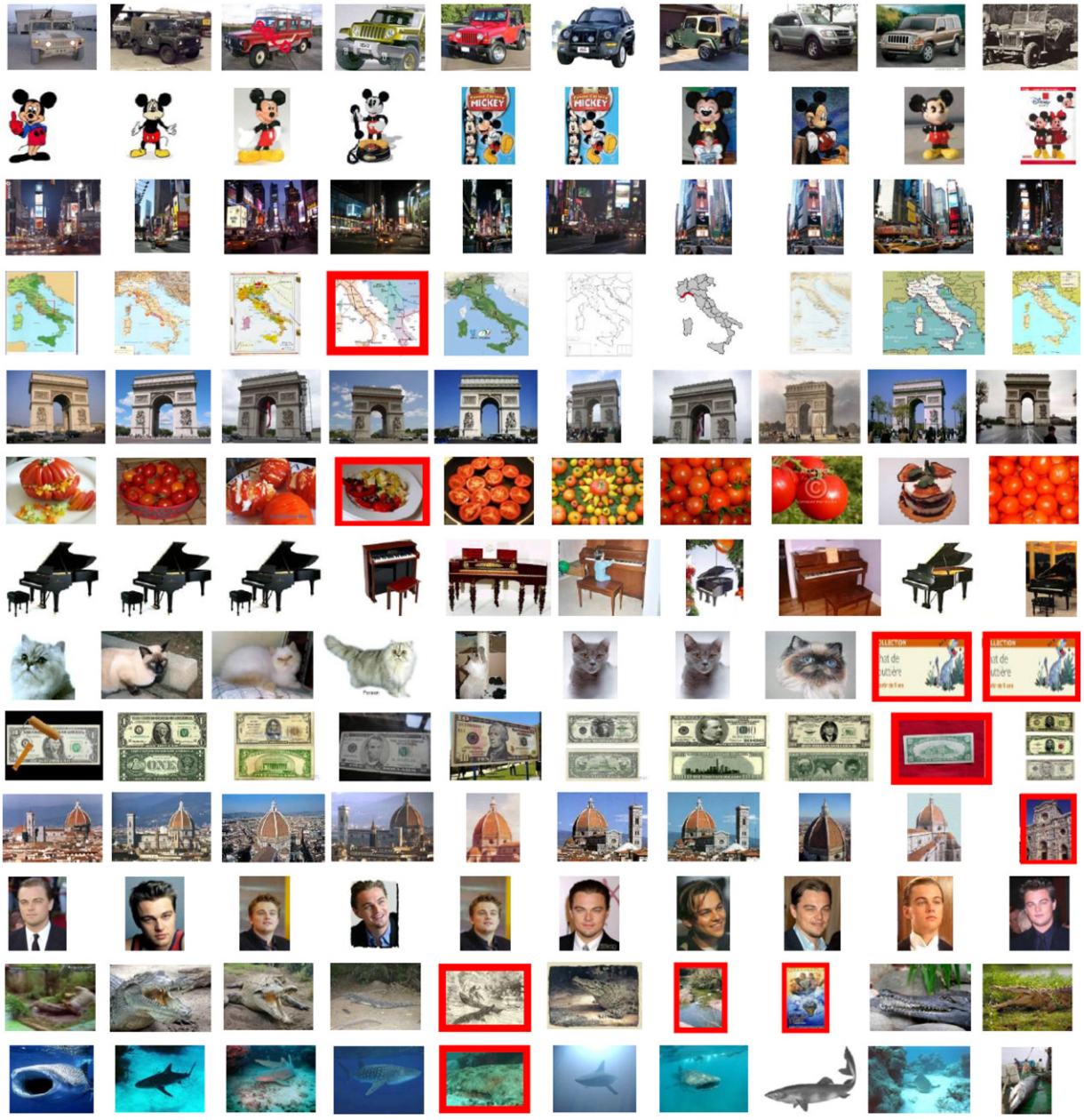

Fig. 14. Examples of the retrieval order obtained by our method. Top 10 images for each query are shown. The queries are (from top to bottom): 4×4 (1st row), Mickey (2nd row), Times Square (3rd row), Italy map (4th row), arc de triomphe (5th row), tomato (6th row), piano (7th row), cat (7th row), dollar (8th row), Dome Florence (9th row), Leonardo di Caprio (10th row), crocodile (11th row), and shark (12th row). The irrelevant images for each query are marked with red.

performs poorly when the ratio of positive instances in the ranking is very small; the AP is especially low when the number of positive instances falls below 3. In this case, the MI classifiers cannot perform well, since there are relatively very few examples to learn from.

We also observe that, for queries that have one or more dominant groups, the performance can be relatively poor. For example, in "Jack Black" query, the dominant set is the black jack table and the multiple instance bags are dominated by such images. Similarly, for "Orsay Museum" query, most of the images show the interior of the museum, whereas only the exterior of the museum is labeled as positive. Our approach tends to rank the interior set of images higher in the retrieval list, and therefore the performance of those queries is inferior. More examples of such cases, where there are more than one dominant group in the query are shown in Fig. 15.

5. Conclusion

In this work, we propose a simple yet effective approach based on Multiple Instance Learning for the problem of image re-ranking. Our approach relies on the construction of multiple candidate MI-bags based on the retrieval order of the images. Assuming that the initial retrieval list contains images of interest, our approach constructs multiple bags and learns multiple MI-classifiers over these bags. Then, the images are re-ranked based on the decision scores of the resulting ensemble of MI-classifiers. Our approach is shown to perform quite successfully compared to the state-of-the-art and significantly outperforms the initial ranking list of produced by the search engines.

Our approach does not make use of any explicit feedback, or auxiliary data such as surrounding text or additional training data. The presented

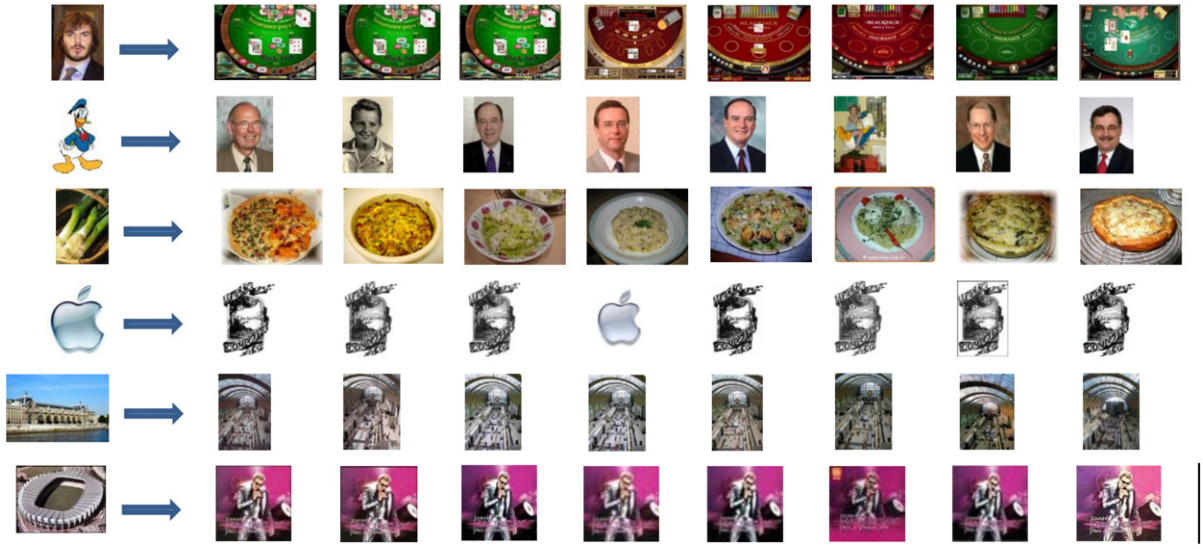

Fig. 15. Examples for the cases in which our method performs relatively poor. For each query, the positive example is given to the left of the list, and to the right is the re-ranked order obtained by our algorithm. The queries are (from top to bottom): Jack Black (1st row), Donald Duck (2nd row), leeks (3rd row), logo apple (4th row), Orsay museum (5th row), and Parc des Princes (6th row). As it can be seen, in these queries, there is more than one dominant visual case in the retrieval list, and our method focuses on the more frequent one. For example, for Orsay museum query, the images returned are mostly from the inside of the museum, which are labeled as negative for that query. Similarly, for the "leek" query, the returned images mostly consist of dishes made with leek, which is also another dominant visual occurrence and also labeled as negative.

method only relies on the visual content of the retrieved images. Given the simplicity of the approach, it can easily be incorporated to more sophisticated schemes, where more complex learning algorithms or more complex visual features are utilized. Considering additional modalities of data can also be explored as a future direction.

Acknowledgments

This work was supported in part by a Google Research Award and the Scientific and Technological Research Council of Turkey (TUBITAK) Career Development Award 112E149.

References

[1] X. Tian, D. Tao, Visual reranking: from objectives to strategies, IEEE Multimed. 18 (2011) 12–21.
[2] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from Google's image search, Proceedings of the 10th International Conference on Computer Vision, (ICCV) Beijing, China, vol. 2, 2005, pp. 1816–1823.
[3] J. Krapac, M. Allan, J. Verbeek, F. Jurie, Improving web image search results using query-relative classifiers, IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10), IEEE Computer Society, San Francisco, United States, 2010, pp. 1094–1101, http://dx.doi.org/10.1109/CVPR.2010.5540092, (URL: http://hal.inria.fr/inria-00548636).
[4] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2008) 1–60.
[5] M. Fritz, B. Schiele, Decomposition, discovery and detection of visual categories using topic models, IEEE Conference on Computer Vision & Pattern Recognition (CVPR), IEEE Computer Society, Anchorage, United States, 2008.
[6] W.H. Hsu, L.S. Kennedy, S.-F. Chang, Reranking methods for visual search, IEEE MultiMed. 14 (2007) 14–22.
[7] Y. Jing, S. Baluja, VisualRank: applying PageRank to large-scale image search, IEEE Trans. Pattern Recognit Mach. Intell. (TPAMI) 30 (2008) 1877–1890.
[8] W. Liu, Y.-G. Jiang, J. Luo, S.-F. Chang, Noise resistant graph ranking for improved web image search, IEEE Conference on Computer Vision & Pattern Recognition (CVPR), IEEE Computer Society, Colorado Springs, United States, 2011.
[9] T.L. Berg, D.A. Forsyth, Animals on the web, CVPR, 2006.
[10] D. Grangier, S. Bengio, A discriminative kernel-based model to rank images from text queries, IEEE Trans. Pattern Recognit. Mach. Intell. (TPAMI) 30 (2008) 1371–1384.
[11] F. Schroff, A. Criminisi, A. Zisserman, Harvesting image databases from the web, ICCV, 2007, pp. 1–8.
[12] B. Geng, L. Yang, C. Xu, X.-S. Hua, Content-aware ranking for visual search, CVPR, 2010.
[13] Y. Chen, J. Bi, J.Z. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE TPAMI 28 (2006) 1931–1947.
[14] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, NIPS, MIT Press, 2003, pp. 561–568.
[15] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, ICML, 1998, pp. 341–349.
[16] P. Viola, J.C. Platt, C. Zhang, Multiple instance boosting for object detection, NIPS, 18, MIT Press, 2006, pp. 1419–1426.
[17] P. Dollár, B. Babenko, S. Belongie, P. Perona, Z. Tu, Multiple component learning for object detection, ECCV, 2008.
[18] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, CVPR, 2009.
[19] B. Zeisl, C. Leistner, A. Saffari, H. Bischof, On-line semi-supervised multiple-instance boosting, Computer Vision and Pattern Recognition (CVPR), 2010.
[20] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, Z. Wang, Joint multi-label multi-instance learning for image classification, CVPR, 2008, pp. 1–8.
[21] O. Yakhnenko, V. Honavar, Multi-instance multi-label learning for image classification with large vocabularies, Proceedings of the British Machine Vision Conference, 2011, pp. 59.1–59.12.
[22] W. Li, L. Duan, D. Xu, I.W. Tsang, Text-based image retrieval using progressive multi-instance learning, ICCV, 2011.
[23] L. Duan, W. Li, I.W. Tsang, D. Xu, Improving web image search by bag-based re-ranking, IEEE Trans. Image Process. (T-IP) (2011) 3280–3290.
[24] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[25] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, http://www.vlfeat.org/ 2008.
[26] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[27] S. Vijayanarasimhan, K. Grauman, Keywords to visual categories: multiple-instance learning for weakly supervised object categorization, CVPR, 2008.
[28] N. Morioka, J. Wang, Robust visual reranking via sparsity and ranking constraints, Proceedings of the 19th ACM international conference on Multimedia, MM '11, 2011, pp. 533–542.
[29] J. Lu, J. Zhou, J. Wang, T. Mei, X.-S. Hua, S. Li, Image search results refinement via outlier detection using deep contexts, Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3029–3036.
[30] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (2010) 1–39.
[31] Y. Han, Y. Yang, X. Zhou, Co-regularized ensemble for feature selection, International Joint Conferences on Artificial Intelligence, (IJCAI), 2013.
[32] W.H. Hsu, L.S. Kennedy, S.-F. Chang, Video search reranking via information bottleneck principle, ACM Multimedia, 2006.
[33] Y. Liu, T. Mei, M. Wang, X. Wu, X.-S. Hua, Typicality-based visual search reranking, IEEE Trans. Circ. Syst. Video Technol. 20 (2010) 749–755.
[34] F. Wu, Y. Han, Q. Tian, Y. Zhuang, Multi-label boosting for image annotation by structural grouping sparsity, ACM Multimedia, 2010. 15–24.
[35] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 723–742.
[36] X. He, W.-Y. Ma, H.-J. Zhang, Learning an image manifold for retrieval, Proc. ACM Int. Conference on Multimedia, 2004, pp. 17–23.
[37] R. Fergus, Y. Weiss, A. Torralba, Semi-supervised learning in gigantic image collections, NIPS, 2009.

[38] V. Jain, M. Varma, Learning to re-rank: query-dependent image re-ranking using click data, Proceedings of the 20th international conference on World wide web (WWW '11), 2011, pp. 277–286.
[39] N. Ben-Haim, B. Babenko, S. Belongie, Improving web-based image search via content based clustering, CVPR Workshop, SLAM, 2006.
[40] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recogn. 45 (2012) 346–362.
[41] J. Li, N.M. Allison, Relevance feedback in content-based image retrieval: a survey, Handb. Neural Inf. Process. 49 (2013) 433–469.